APLICACIONS ESTADÍSTIQUES AMB FRAMEWORK

Josep Miquel Bujosa Bravo

Santiago Manrique Catalán

Abril 1991

Títol: **Aplicacions estadístiques amb Framework**

Autors: **Josep Miquel Bujosa Bravo i Santiago Manrique Catalán**

Generalitat de Catalunya
Departament d'Ensenyament
Programa d'Informàtica Educativa

ÍNDEX

1 INTRODUCCIÓ 4

2 UNIVAR 5

2.1 OBJECTIUS 5

3 CORREG 7

3.1 OBJECTIUS 7

4 SONDEIG 9

4.1 OBJECTIUS 9

5 PRÀCTIQUES 11

5.1 Pràctiques amb UNIVAR 11

5.1.1 Pràctica 1 11
5.1.2 Pràctica 2 12
5.1.3 Pràctica 3 13

5.2 Pràctiques amb CORREG 15

5.2.1 Pràctica 1 15
5.2.2 Pràctica 2 17
5.2.3 Pràctica 3 19
5.2.4 Pràctica 4 20
5.2.5 Pràctica 5 21

5.3 Pràctiques amb SONDEIG 22

5.3.1 Pràctica 1 22
5.3.2 Pràctica 2 23

6 ANNEX 24

7 BIBLIOGRAFIA 25

INTRODUCCIÓ

En aquestes pàgines volem documentar les aplicacions UNIVAR, CORREG i SONDEIG incloses en el disc anomenat *Aplicacions estadístiques en Framework* (fitxers UNIVAR.FW3, CORREG.FW3 i SONDEIG.FW3), dissenyades bàsicament amb fulls de càlcul i pensades com a suport a l'hora de tractar a classe certs aspectes de probabilitat i estadística.

Per una bona utilització d'aquestes aplicacions s'haurà de treballar amb el Framework configurat en mode gràfic, ja que en cas contrari no s'arribarien a veure correctament els gràfics sense fer un zoom, ni tampoc s'apreciarien els exponents i subíndexs de determinades expressions que hi intervenen.

En aquest document, trobareu, en primer lloc, una descripció breu de les tres aplicacions i, a continuació, unes pràctiques que poden acabar d'aclarir el respectiu funcionament. En aquestes pràctiques es fa referència a uns fitxers de dades que estan al disc esmentat anteriorment.

Al final, hem inclòs en un annex algunes fórmules poc usades en un nivell no universitari però que, malgrat això, hem utilitzat per a determinats càlculs.

UNIVAR

```
[UNIVAR]
[Dades inicials]
                                               Entrada de dades
     Base de dades:                        Només dades  Dades-freqüències
    Nom camp dades:           Tractament:
Nom camp freqüèn.:                 Discret       1            2
Condició (filtre):
Cas escollit(1-4):                 Continu       3            4

[Escala]                              [Opcions]
            Valor mínim:
            Valor màxim:
                                         Tabulació i Histograma   x
        Extrem inferior:                            Estadístics   x
        Extrem superior:               Diagrama centralització   x
        Long. intervals:                    Diagrames de caixa   x
     Nombre d'intervals:
```

OBJECTIUS

L'objectiu fonamental d'aquesta aplicació és l'anàlisi i tractament estadístics de dades univariants.

Aquestes han d'estar situades en una base de dades, de manera que cada camp representi una variable. L'aplicació recull els valors numèrics del camp de la base de dades indicat i passa a fer els càlculs necessaris. L'usuari pot escollir entre considerar els valors un a un, de manera que la freqüència absoluta d'un determinat valor vindrà indicada per la seva repetició, o bé amb la modalitat de dada-freqüència. En aquest cas, s'haurà d'indicar, a més del camp de la variable, a quin camp de la mateixa base es troben les freqüències absolutes de cada valor. També es pot escollir un camp discriminatori, de manera que segons els valors d'aquest camp es consideri o no la dada del mateix registre. La indicació d'aquest camp s'ha de fer de la manera habitual com s'escriuen els filtres en les bases de dades de Framework, és a dir:

Nom de camp = valor

A més d'aquestes possibilitats esmentades en el format de l'entrada de dades, l'usuari pot escollir entre uns tractaments discret o continu d'aquestes dades. La diferència fonamental entre aquests dos tractaments està en què en el discret cadascuna

de les dades introduïdes és considerada pel càlcul dels diferents paràmetres estadístics, mentre que en el tractament continu és la marca de classe el valor que entra en joc pels càlculs.

L'usuari també pot escollir la quantitat, les mides i l'extrem inferior dels intervals que serviran per una tabulació posterior.

L'aplicació ofereix quatre opcions per escollir: *Tabulació i histograma*, *estadístics*, *diagrama de centralització* i *diagrames de caixa*. Es poden escollir les quatre o algunes d'elles, segons l'interès de la sessió de treball.

Amb l'opció *tabulació i histograma*, l'aplicació mostrarà els intervals de classe segons l'indicat a la finestra "escala", i les freqüències absolutes respectives. A més, dibuixarà un gràfic de barres, a partir de la classificació feta.

Si s'escull l'opció *estadístics*, l'aplicació mostrarà el nombre de dades que s'estan tractant, la mitjana, el màxim valor tractat, el mínim, el rang, la mediana, el primer quartil, el tercer quartil, la distància interquartílica, la variància, la desviació estàndard, el coeficient de variació i el percentatge de dades incloses dins de l'interval de la mitjana ± una desviació estàndard i dins del de la mitjana ± dues desviacions estàndard. Cal aclarir que tant la variància com la desviació estàndard estan calculades fent servir n com a divisor de la fórmula, i no n-1 com és el cas de les funcions de FRED @VAR i @STD.

En triar les opcions de *diagrama de centralització* i *diagrames de caixa*, l'aplicació dibuixarà uns gràfics representant, en el primer cas, la situació de la mitjana i de la mediana en la distribució de valors estudiats, i, en el segon cas, la situació relativa dels quartils, la mediana inclosa, en tota de la distribució. En el primer diagrama es podrà observar la diferència entre la mediana i de la mitjana com a paràmetres de centralització. En el segon diagrama, en canvi, es poden observar les possibles acumulacions de valors al centre o cap als extrems, la qual cosa ens permet deduir la forma de la distribució.

Tots els gràfics es presenten sense cap fórmula al marc de les finestres respectives. Opcionalment, se'n poden fer còpies, amb la tecla F8, per a possibles comparacions posteriors.

CORREG

```
[CORREG]
  [Dades inicials]
        Base de dades:                (Nom de la base de dades)
           Variable x:                (Nom del camp x)
           Variable y:                (Nom del camp y)
    Condició (filtre):                (Sense filtre agafa tots els valors)

 Eix X:                                   Títol general:
       Rètol:                       Mida dels punts(1-5): 3
       Mínim:                          Cas escollit(0-6): 0
       Màxim:
   Increment:                             Models          1 2 3 4 5 6 0

 Eix Y:                                     Lineal y(x):  ♦ ♦ ♦ ♦
       Rètol:                               Lineal x(y):    ♦
       Mínim:                                 Quadràtic:      ♦   ♦
       Màxim:                               Exponencial:        ♦   ♦
   Increment:                     (0=només núvol de punts)
```

OBJECTIUS

L'objectiu d'aquesta aplicació és l'estudi de la correlació lineal entre dues variables numèriques, així com l'obtenció de models de regressió lineal, quadràtica i exponencial, tot analitzant aquell que s'ajusta millor a les dades i amb la possibilitat de fer prediccions, seguint els models escollits.

El funcionament és molt similar al de l'aplicació UNIVAR, explicada anteriorment. En aquest cas, també agafarà els valors de les variabls d'una base de dades on cada camp n'és una. S'haurà d'indicar el nom de la base, el nom del camp considerat com a variable X, i el considerat com a variable Y. Per analitzar només un subconjunt de totes les dades, hi ha la possibilitat d'escriure una condició en forma de filtre, fent referència a un camp determinat que serveixi per discriminar les dades considerades. Si es deixa en blanc la cel·la corresponent a la condició esmentada, l'aplicació considerarà tots els parells de dades numèriques dels camp indicats. En qualsevol cas, aquells registres en què falti alguna dada en aquests camps, no seran considerats.

De cara a la representació gràfica, es poden escollir els rètols que sortiran a cada eix ,els valors mínim i màxim per a cada variable ,els increments que serviran per fer les

separacions als eixos, la mida dels punts i el títol general del gràfic. Per defecte, l'aplicació dóna uns valors per a tots aquests detalls, que sempre es poden modificar.

Per estudiar la regressió, es poden escollir set cassos diferents (del 0 al 6). Si s'escull el cas 0, no es calcularà cap model de regressió i, des del punt de vista gràfic, només es representarà el núvol de punts de les dues variables. En el cas 1, es calcularà i es representarà la recta de regressió que resulta de cosiderar la variable Y com dependent i la X com independent (Y(X)). En el cas 2, es farà el mateix amb, a més de la recta de regressió abans comentada, la resultant de considerar la Y com a variable independent i la X com a variable dependent (X(Y)). En el cas 3, es calcularan i representaran la recta de regressió Y(X) i la paràbola de regressió. En el cas 4, la recta Y(X) i la funció exponencial com a model de regressió. Els cassos 5 i 6 estan pensats pels models quadràtics (paràbola) i exponencial respectivament, considerats de manera individual. En aquells casos en què s'estudien dos models a l'hora (2, 3 i 4), la representació gràfica es fa superposant en uns mateixos eixos els dos models, a fi de poder-los comparar millor.

Sigui quin sigui el cas escollit, l'aplicació calcularà el nombre de parells de dades considerades, la mitjana per a cada variable, les desviacions estàndard Sx i Sy, la covariància Sxy i el coeficient de correlació lineal (de Pearson) r. Segons els models triats, mostrarà en pantalla els paràmetres que els determinen, així com el coeficient de determinació R^2 per a cadascun. Aquest coeficent, els valors del qual estan sempre entre 0 i 1, ens indicarà quin model s'ajusta més al núvol de punts, tot observant el grau d'aproximació a 1 (per a més aclariments sobre la fórmula veieu l'annex).

Com en l'aplicació anterior, els gràfics estan lliures de fórmules als marcs de les seves finestres i poden ser copiats amb F8 fora del marc més extern de l'aplicació.

L'usuari pot cercar imatges segons els models triats, introduïnt els valors per a la variable X en el full anomenat prediccions.

SONDEIG

```
[SONDEIG]

        Mida de la població:  10000000   (Si és gran, no té importància.)

     Nivell de confiança(%):   95,50     (90, 95, 95.5[2σ], 99 o 99.7[3σ])

 Opcions:    1. Entrar la mida de la mostra per estimar els errors
             2. Entrar l'error màxim per estimar la mida de la mostra

   Escolliu una opció (1-2):    1

          Mida de la mostra:   800       Error màxim (p=q=50%)= ±3,54 %

     Percentatge mostral(%):   24,30     %  Error= ±3,03 %

   Interval de confiança(%):  (21,27 , 27,33)
```

OBJECTIUS

Aquesta aplicació està pensada per facilitar l'estimació de la proporció d'elements d'una població amb una característica comú, a partir de la proporció obtinguda d'una mostra ben escollida de la mateixa població. Facilitarà, per tant, la comprensió de les anomenades "fitxes tècniques" de qualsevol enquesta d'opinió, que tan sovint apareixen en els mitjants de comunicació, i permetrà conèixer, abans de fer qualsevol enquesta, quina és la mida de la mostra aconsellable si no es vol passar d'un error determinat.

L'usuari haurà d'entrar la mida de la població subjecta a estudi i haurà d'escollir el nivell de confiança desitjat d'entre els cinc possibles que s'ofereixen. La mida de la població tindrà molta importància per a poblacions petites, ja que se segueix el model d'extracció sense reposició (és a dir, que a una mateixa persona no se la consulta dues vegades per a la mateixa qüestió).

Es poden escollir dues opcions:

- En escollir l'opció 1, s'ha d'entrar la mida de la mostra i la proporció (percentatge) obtinguda a partir d'aquesta. L'aplicació retorna l'error màxim comés (per a p=q=50%) i l'error ajustat a aquest percentatge mostral, així com el

correspondent interval de confiança. Tots aquests resultats seran importants a l'hora d'extrapolar la proporció mostral a la població.

-Si es tria l'opció 2, s'ha d'entrar el màxim error percentual que un està disposat a cometre. L'aplicació retornarà la mida de la mostra necessària per a no superar aquest error.

Cal tenir present que en tota l'aplicació s'està fent servir l'aproximació de la binomial per la normal. Això implica que si n és la mida de la mostra i p la proporció poblacional, quan np i n(1-p) siguin "petits" (≤ 5), els resultats poden no tenir una bona precisió. Per a més informació sobre les fórmules utilitzades veieu l'annex.

```
[SONDEIG]
        Mida de la població:  10000000   (Si és gran, no té importància.)

     Nivell de confiança(%):   95,50     (90, 95, 95.5[2σ], 99 o 99.7[3σ])

Opcions:     1. Entrar la mida de la mostra per estimar els errors
             2. Entrar l'error màxim per estimar la mida de la mostra

    Escolliu una opció (1-2):     2

     Error màxim (p=q=50%): ±     5         %


Mida aconsellada de la mostra = 400
```

PRÀCTIQUES

Pràctiques amb UNIVAR

Pràctica 1

-Recupereu del disc els fitxers UNIVAR.FW3 i ALUMNES.FW3. Aquest últim és una base de dades que conté un conjunt de valors antropomètrics de nois i noies de tercer de BUP. Començarem a analitzar la variable Pes. Fixeu-vos que hi ha un camp anomenat Sexe on figura un 1 si es tracta d'un noi, i un 2 si es tracta d'una noia.

-Premeu **F5** des del marc de la finestra UNIVAR. El cursor s'haurà situat a l'interior del full Dades inicials.

-Entreu a la cel·la B2 el nom de la base de dades: Alumnes.

-Entreu a la cel·la B3 el nom del camp on tenim els pesos: Pes.

-Deixeu en blanc les dues cel·les següents i passeu a la B6. Aquí cal escollir el cas. Observant el quadre de la dreta, es veu que es disposa de quatre casos, segons la forma d'entrar les dades i el tractament que se'n faci d'elles. Escolliu el cas 1, és a dir, les dades del camp triat (Pes) es consideren sense anar a cercar un altre camp on figurin les freqüències absolutes. Aquestes dades tindran un tractament discret (Més endavant veureu la diferència entre un tractament discret i un tractament continu).

-En haver entrat les referències del paràgraf anterior, premeu **F5**. El cursor ha quedat situat a l'interior del full Escala, on trobareu els valors màxim i mínim de les dades, els extrems inferior i superior considerats per fer la tabulació, així com la quantitat d'intervals que es construiran i la seva mida. L'extrem inferior, el nombre d'intervals i la seva longitud son modificables. En funció dels valors d'aquests, es calcularà el valor de l'extrem superior. Deixeu 40 com a extrem inferior, 6 com a longitud dels intervals i 7 com a nombre d'intervals. Premeu **F5**.

-El cursor ha quedat, ara, al full de les opcions. En aquesta primera pràctica només interessa veure la tabulació i l'histograma. Deixeu una **x** a l'apartat corresponent i en blanc les altres opcions. Premeu **F5**.

-Per la pantalla van sortint els diferents intervals, el nombre de valors que pertanyen a cadascun i el gràfic associat.

> -Premeu ↑ per tornar a la pàgina anterior. Situeu el cursor a l'interior del full Escala i entreu com a nova longitud 5 i com a nombre d'intervals 8. Premeu **F5** dos cops. Observeu la nova tabulació que s'ha calculat.

-Conserveu una còpia d'aquesta tabulació nova:

-Amb el cursor al marc de la finestra Tabulació, premeu **F8.**

-Premeu - per sortir fora d'UNIVAR.

-Premeu ↵ per acabar la còpia.

-Anomeneu aquesta finestra copiada amb el nom de Pestots.

Repetireu, ara, el mateix procés però distingint les dades masculines de les femenines.

-Premeu **F5** des del marc de la finestra UNIVAR.

-Afegiu al full Dades inicials la condició o elfiltre sexe=1. D'aquesta manera es farà un filtratge dels registres de la base Alumnes que tenen un 1 al camp Sexe. Amb això s'aconsegueix que només es tinguin en compte les dades dels nois. Premeu **F5.**

-Modifiqueu les dades del full Escala de manera que l'extrem inferior sigui 40, la longitud 5 i el nombre d'intervals sigui 8. Premeu **F5** dos cops.

-Feu una còpia de la finestra Tabulació relativa a les dades dels nois i anomeneu-la Pesnois.

-Repetiu tot el procés anterior posant com a condició (filtre) al full Dades inicials Sexe=2. Així quedaran seleccionades només les dades referides a les noies. Al full Escala, entreu 40 com a extrem inferior, 5 com a longitud dels intervals i 8 com a nombre d'intervals.

-Feu una còpia de la finestra tabulació en el cas de les noies i anomeneu-la Pes noies.

-Compareu les tres finestres obtingudes, col·locant-les en l'origen de la pantalla amb la mateixa mida. Passeu d'una a l'altra amb les tecles de moviment del cursor.

Pràctica 2

L'objectiu d'aquesta pràctica és comparar els pesos dels nois i de les noies de la base de dades ALUMNES, a partir del càlcul dels diferents paràmetres estadístics que permet l'aplicació UNIVAR.

-Heu de tenir carregats els fitxers UNIVAR.FW3 i ALUMNES.FW3.

-Premeu **F5** des del marc de la finestra UNIVAR.

-Al full dades inicials han d'aparèixer el nom Alumnes, el camp Pes, la condició Sexe=1 (per als nois) i el cas 1.

-Des del full Escala podeu prémer **F5** sense modificar res.

-Al full Opcions suprimiu la x de l'opció Tabulació i Histograma i poseu-les a les altres opcions. Premeu **F5**.

-Apareix en pantalla una nova pàgina on hi trobareu diferents paràmetres estadístics i uns gràfics explicatius de la distribució. Analitzeu-los.

-Interessa fer una còpia de la finestra Distribució per a futures comparacions. Procediu de la manera habitual. Anomeneu NOIS a la còpia nova.

-Premeu **F5** des del marc del full UNIVAR.

-Repetiu tot el procés anterior per a les noies, entrant Sexe=2 com a condició al full Dades inicials.

-Feu una còpia de la finestra Distribució creada a partir de les dades de les noies i anomeneu-la NOIES.

-Compareu les dues finestres creades.

Opcional: repetiu els estudis fets en aquestes dues pràctiques per a altres variables de la base Alumnes.

Pràctica 3

L'objectiu d'aquesta pràctica és fer servir l'aplicació UNIVAR quan les dades que volgueu tractar es trobin en forma de dada-freqüència.

-Recupereu el fitxer SOUS.FW3. Aquesta petita base de dades conté dos camps. En el primer, anomenat Sou, trobareu els diferents sous, expressats en milers de pessetes, dels treballadors d'una suposada empresa. A l'altre camp, Fsou, hi figuren el nombre de treballadors que gaudeixen del sou corresponent. Són les freqüències absolutes de la variable Sou. Així, per exemple hi ha setze persones que cobren 140.000 pessetes, i només dues amb 220.000 pessetes.

-Premeu **F5** des del marc de la finestra UNIVAR.

-El nom de la base de dades és Sous. El nom del camp de dades és Sou. Cal entrar el nom del camp de freqüències Fsou a la cel·la corresponent. No hi ha d'haver cap filtre. Trieu el cas 2 (tractament discret amb entrada de dades-freqüències). Després d'actualitzar el full Dades inicials premeu **F5**.

-No modifiqueu cap valor del full Escala i premeu **F5**.

-Marqueu amb una **x** totes les opcions del full Opcions i premeu **F5**.

-Apareix la pàgina de tabulacions. A més del que es veu en pantalla, també s'estan calculant tots els paràmetres necessaris per satisfer les opcions escollides. Quan s'hagin acabats els càlculs, podreu prémer ↓ des del marc de Tabulació i veureu els estadístics i els diagrames.

-Fixeu-vos en els diferents paràmetres estadístics. Quin paràmetre és més representatiu per descriure el sou que es cobra en aquesta empresa, la mitjana o la mediana?

-Per tornar a veure les tabulacions haureu de prémer ↑ des del marc de Distribució.

-Modifiqueu alguns aspectes del full Escala.

-Col·loqueu el cursor a l'interior del full Escala.

-Deixeu 100 com a extrem inferior. Entreu 21 com a longitud dels intervals i 6 com a nombre d'intervals.

-Premeu **F5** dos cops.

-Analitzeu el resultat obtingut.

Opcional: repetiu el procés amb les mateixes dades escollint el cas 4, és a dir, considerant un tractament continu. Observeu que tots els estadístics es calculen utilitzant les marques de classe de cada interval fixat a partir del full Escala, tret dels quartils, mediana inclosa, ja que, per aquests, es consideren les dades uniformement repartides en cada interval.

Pràctiques amb CORREG

Pràctica 1

L'objectiu d'aquesta pràctica és estudiar les possibles correlacions lineals existents entre unes variables obtingudes a partir d'unes proves d'Educació Física fetes per nois de primer de BUP.

-Recupereu del disc els fitxers CORREG.FW3 i EF.FW3. Aquest últim és una base de dades que conté les variables amb què treballareu. Cada registre representa els resultats obtinguts per un alumne determinat en cada prova, així com el grup-classe al qual pertany. Analitzem els camps:

GRUP: representa els diferents grups-classe dels alumnes.

ABDOM: conté el nombre de flexions abdominals que ha realitzat cada alumne en un temps determinat.

POTCA: representa els centímetres de llargada que ha saltat cada noi sense donar-se impuls i amb els peus junts. Ens dóna una mesura de la potència de cames.

TEMPS: indica el temps, en segons, obtingut per cada individu, en recórrer una distància de 60 metres llisos.

VELOM: expressa en metres/segon la velocitat mitjana de cada participant, calculada a partir de les dades de la variable TEMPS.

Es tracta d'estudiar les possibles correlacions que hi poden haver entre aquestes variables.

Fase 1

Comenceu per les variables POTCA i VELOM. Esbrineu quina relació hi ha entre la potència de cames i la velocitat mitjana en una prova de 60 metres.

-Premeu **F5** des del marc de la finestra de CORREG. El cursor ha quedat situat a l'interior del full Dades inicials, on heu d'escriure a les cel·les B1, B2 i B3 (mireu la línia d'estat) el nom de la base de dades EF, la variable x POTCA i la variable y VELOM, respectivament. La cel·la B4 ha de quedar buida, ja que no imposem cap condició i volem considerar tots els valors de les dues variables seleccionades.

-Premeu un altre cop **F5** des de l'interior de Dades inicials. Ara el cursor ha passat al full Paràmetres. A la part esquerra d'aquest full ha de sortir una sèrie de paràmetres per a cada eix, els quals són necessaris per fer el gràfic del núvol de punts. Aquests paràmetres es poden variar. En aquest cas només

modificareu els títols de cada eix. Escriviu Potència de cames en comptes de POTCA (cel·la B2 de Paràmetres) i Velocitat mitjana en comptes de VELOM (cel·la B8).

A la part superior dreta escriviu el títol general Correlació (cel·la D1). A la cel·la D2 podeu escollir la mida dels punts. A la cel·la D3 hi ha d'haver un 0, ja que només voleu veure el núvol de punts. Si volguéssiu fer una regressió, haurieu de triar el cas segons els mòdels possibles indicats a l'esquema de la part inferior.

-Premeu **F5** en acabar d'escriure els paràmetres desitjats . Heu passat a la pantalla següent, a la qual hi ha tres finestres. A la finestra Càlculs apareixen el nombre d'alumnes, les mitjanes de cada variable, les seves desviacions estàndard, la covariància S_{xy} i el coeficient de correlació lineal r. Aquest coeficient, els valors del qual estan sempre entre -1 i 1, indica si la relació entre les dues variables tractades és "més o menys" lineal. Això significa que per a valors propers a 1 o a -1, la relació és "molt" líneal, creixent o decreixent respectivament, és a dir, que la gràfica s'assemblarà a una recta (creixent o decreixent, segons el signe). Com més proper estigui a zero, el gràfic s'assemblarà menys a una recta. En aquest cas ha sortit un valor possitiu i proper a 1. Això significa en general, que com més potència de cames tingui una persona, més ràpid pot córrer.

-Premeu **F5** per veure la gràfica , tal com indica el rètol. Observeu-la i comproveu les afirmacions fetes anteriorment. Fixeu-vos també amb els rètols. Són els escrits al full Paràmetres.

-Feu una còpia d'aquesta gràfica per fer comparacions :

-Siteu el cursor al marc de la finestra Gràfica.

-Premeu **F8**.

-Situeu el cursor al marc de CORREG.

-Premeu ↵ .

-Canvieu el nom de la còpia feta. Abans haureu d'haver seleccionat Permetre edició del menú Edit. Anomeneu-la POTCAVEL (per exemple).

Opcional: repetiu el procés amb POTCA i TEMPS. Sortirà un valor de r proper a -1.

Fase 2

Compareu els valors de les variables ABDOM (nombre de flexions abdominals) i VELOM (velocitat mitjana).

-Repetiu tot el procés anterior i adapteu-lo a les variables esmentades abans.

-En aquest cas el valor de r és proper a zero. Si observeu el gràfic podreu comprovar el que significa això, i podreu treure conclusions respecte a la relació que hi ha entre les proves físiques considerades. Cal recordar que un valor proper a zero del coeficient r, no sempre significa que no hi hagi cap relació entre les variables sinó que la relació, si existeix, no té un comportament similar a una recta.

-Feu una còpia del gràfic obtingut i compareu-lo amb l'anterior.

Opcional: esbrineu d'altres relacions possibles entre les variables d'EF.

<u>Pràctica 2</u>

L'objectiu d'aquesta pràctica és esbrinar la correlació lineal entre unes variables de dades antropomètriques, així com veure la possibilitat d'analitzar subconjunts de dades de les mateixes variables atanent a una característica classificatòria, com pot ser el sexe dels individus. Concretament, analitzeu les dades del fitxer TERCER.FW3, fent servir l'aplicació CORREG.FW3.

Fase 1

-Recupereu del disc el fitxer TERCER.FW3. Aquest fitxer és una base de dades amb set camps i conté un conjunt de mides preses a nois i noies de tercer de BUP. En el camp anomenat SEXE, s'indica amb un 1 el registre d'un noi, i amb un 2, el d'una noia. Com podeu observar, alguns camps no tenen totes les dades. L'aplicació CORREG, en analitzar aquestes variables, ignorarà els individus amb alguna dada desconeguda entre les tractades. Comenceu per analitzar la correlació entre les variables ALÇADA i PUNTESMANS (distància en centímetres des d'una punta d'una mà a l'altra de l'altra ma, amb els braços estirats formant una creu).

-Premeu **F5** des del marc de CORREG.

-Escriviu el nom de la base de dades TERCER. La variable **x** serà ALÇADA i la variable y PUNTESMANS. Manteniu en blanc la cel·la B4 destinada a un possible filtre.

-Premeu **F5**. Apunteu en un paper les dades que han sortit al full Paràmetres referides al Mínim, al Màxim i a l'increment de cada eix. Posteriorment les farem servir per comparar gràfics. Escriviu com a títol general "Nois i noies" (cel·la D1). Fixeu la mida dels punts i trieu el cas 0.

-Premeu **F5**. Observeu els resultats del full Càlculs. Ha considerat els 59 alumnes ja que no faltava cap dada en aquestes variables, i ha calculat diferents paràmetres. Fixeu-vos en l'alt coeficient de correlació lineal. Què significa això? Anoteu aquest valor de r per a comparacions futures.

-Premeu **F5** per veure la gràfica (pot trigar una mica).

-Feu una còpia de la gràfica a l'exterior de CORREG i anomeneu-la TOTS, seguint les indicacions de la pràctica anterior. Fixeu-vos en la forma del núvol de punts.

A continuació interessa fer al mateix analitzant els nois i les noies per separat, per poder fer comparacions posteriors.

-Premeu **F5** des del marc CORREG. Manteniu els noms de la base de dades, de la variable x i de la variable y. Escriviu a la cel·la B4 la condició sexe=1. D'aquesta manera, de tota la base es filtraran només les dades dels nois.

-Premeu **F5**. Algunes dades referides als eixos s'han modificat. Això és degut al fet que CORREG busca una escala òptima per a les dades que ara té. En aquest cas, però, interessa mantenir la mateixa escala que abans, per poder comparar millor els gràfics. Varieu les dades de manera que quedin com en el cas anterior. Escriviu com a títol general "Nois". Manteniu la mateixa mida dels punts i el cas 0.

-Premeu **F5**. Observeu el canvi dels càlculs. El valor de r s'ha modificat però, tot i així, és bastant alt. Anoteu-lo.

-Premeu **F5** per veure la gràfica. Feu una còpia a l'exterior de CORREG. Anomeneu-la Nois.

-Repetiu tot el procés anterior adaptant-lo a les noies. La condició del filtre ha de ser sexe=2. No us oblideu d'escriure les mateixes dades per a l'escala dels eixos que les del cas general. Feu una còpia del nou gràfic i anomeneu-la Noies.

-Creeu una finestra buida de text i anomeneu-la ALÇADA-PUNTESMANS.

-Modifiqueu les seves mides de manera que ocupi tota la pantalla (**F4 F4** ↵).

-Traslladeu els tres gràfics (Tots, Nois i Noies) a l'interior d'aquesta finestra. Activeu l'opció <u>Liure desplaçament</u> del menú <u>Edit</u>.

-Situeu les tres finestres gràfiques, una sobre l'altra, al cantó superior esquerre de la finestra contenidora. Dimensioneu-les al màxim.

-Premeu successivament qualsevol tecla de desplaçament del cursor i aconseguireu comparar ràpidament els diferents núvols de punts.

Fase 2

En aquesta part heu de repetir tot el procés anterior amb les variables NUMSAB (número de peu) i PUNY (mida del puny). La novetat d'aquesta pràctica consisteix en veure que, tot i que surt un valor de r acceptable considerant tots els alumnes, en considerar per separat els nois i les noies, surten valors molt més propers a zero. Aquesta paradoxa inicial s'explica comparant els tres gràfics obtinguts, i dóna una veu d'alerta davant les conclusions massa precipitades.

Opcional: esbrineu més correlacions amb d'altres variables de la base TERCER.

Pràctica 3

En aquesta pràctica es presenta el concepte de regressió lineal, a partir de l'anàlisi de la relació entre les alçades d'uns pares i els seus fills. En un estudi similar, Francis Galton (1822-1911) a finals del segle passat va utilitzar per primer cop el terme "regressió". Fareu servir, a més de l'aplicació CORREG.FW3, el fitxer de dades GALTON.FW3.

Fase 1

-Recupereu del disc el fitxer GALTON.FW3. És una base de dades que conté les alçades mitjanes de 29 pares i les dels seus fills.

-Premeu **F5** des del marc CORREG.

-El nom de la base serà Galton. La variable x serà Pares i la y serà Fills. Deixeu en blanc la cel·la dels filtres.

-Premeu **F5**. Entreu un 1 a la cel·la D3 per escollir el cas de regressió lineal y(x).

-Fixeu-vos en les mitjanes i les desviacions estàndard. En mitjana, els fills són més alts que els pares i tenen una desviació estàndard molt similar. El coeficient de correlació és bastant alt. El coeficient R^2 ens indica quina proporció de variació de la variable y (alçada dels fills) s'explica per la variació de la variable x (alçada dels pares), seguint el model de la recta de regressió y=ax+b, els coeficients de la qual apareixen també en aquest full.

-Premeu **F5** des del marc de Gràfica. Aquesta recta que apareix és l'anomenada recta de regressió. Passa pel punt mitjà del núvol de punts i la considerarem a l'hora de fer prediccions.

-Situeu el cursor al marc de Predicions i premeu **F5**. Entreu un valor per a la x (alçada pares) i calculeu el valor corresponent a la y (alçada fills) seguint la recta de regressió abans calculada.

-Esbrineu quines alçades dels fills poden esperar pares molt baixos (per exemple 155), propers a la mitjana (per exemple 171), o molt alts (per exemple 190). Si us hi fixeu, els pares molt baixos tendeixen a tenir, en mitjana, fills molt més alts que ells; mentre que pares molt alts, segons aquest model, tindran fills no tan alts com ells. Aquest fenomen es coneix com "efecte regressió" i va ser descrit per Galton, comentant que "representa una **regressió** a la mediocritat".

Fase 2

-Premeu **F5** des del marc de CORREG.

-Repetiu tot el procés de la fase anterior, però escollint-ne el cas 2.

-Al full Càlculs apareixen les expressions de dues rectes de regressió: y(x), ja vista anteriorment i x(y), on la variable dependent és l'alçada dels pares (x) i la independent la dels fills (y). Si feu la gràfica les veureu representades totes dues.

-Aneu fent prediccions i observeu els resultats obtinguts.

<u>Pràctica 4</u>

En aquesta pràctica es veurà com la recta de regressió no sempre és el millor model per representar la correlació entre dues variables. Fareu servir els fitxers de dades A.FW3, B.FW3 i C.FW3.

-Recupereu del disc els fitxers A.FW3, B.FW3 i C.FW3. Tots tres contenen dues variables: x i y.

-Executeu l'aplicació CORREG (**F5** des del marc). El nom de la base de dades serà A, les variables x i y. Trieu el model lineal y(x), és a dir, el cas 1.

-Anoteu el coeficient r i l'expressió de la recta de regressió.

-Veieu la gràfica i feu-ne una còpia per a comparacions posteriors.

-Repetiu el procés per a les bases B i C. Com podeu observar, els coeficients de correlació lineal r i les rectes de regressió són pràcticament idèntics en tots tres casos, tot i que els gràfics us demostren que esteu davant de tres exemples molt diferents, on la recta no sempre s'ajusta bé. Observant l'exemple B, es pot intuir que una paràbola s'ajustaria millor. Comproveu això:

-Premeu **F5** des del marc de CORREG. La base de dades és B.

-Escolliu el cas 3. Se superposaran el model lineal (y(x)) i el model quadràtic (paràbola).

-En el full Càlculs apareixeran les expressions de la recta de regressió i els de la paràbola de regressió. Compareu els dos valors d'R^2 . El valor més gran ens indica que el model s'ajusta millor. Val a dir que $0 \le R^2 \le 1$.

-Feu la gràfica per comprovar-ho.

Amb aquests exemples es pot veure com el núvol de punts pot ser de gran utilitat a l'hora de triar el model adequat per fer una regressió.

Pràctica 5

En aquesta pràctica es veurà el model exponencial de regressió. Fareu servir la base de dades LABOR.FW3, la qual conté unes suposades dades obtingudes en un laboratori. A partir d'aquestes dades, es vol ajustar una funció per explicar el fenomen.

-Recupereu del disc el fitxer LABOR.FW3.

-Executeu CORREG. La base de dades és LABOR. Les variables són A i B, respectivament.

-Escolliu el cas 3, és a dir, el model lineal i quadràtic.

-Observeu els resultats de Càlculs. Sembla clar que la paràbola s'ajusta millor a les dades que no pas la recta. Anoteu els diferents valors d'R^2.

-Comproveu-ho fent la gràfica. Feu-ne una còpia.

-Torneu a repetir el procés anterior, però ara escolliu el cas 6, que representa el model exponencial.

-Observeu que al full Càlculs apareixen els coeficients a i b de la funció $y = a \cdot b^x$, i c per a l'equivalent $y = a \cdot e^{cx}$. Fixeu-vos que el valor d'R^2 és més gran que els anteriors.

-Feu la gràfica i , amb una còpia, compareu-la amb l'anterior.

Pràctiques amb SONDEIG

	OTR/IS El Periódico	Instituto Opina La Vanguardia	Demoscopia El País	Resultats electorals
CiU	46,40	47,90	46,50	46,19
PSC	29,20	28,20	27,60	27,52
ERC	5,60	6,50	6,20	7,96
IC	7,70	7	7,10	6,51
PP	6,30	5,60	6,30	5,96
CDS	1,90	1,50	1,10	,91
N.Confiança	95,50	95	95,50	
Mostra	2000	2004	1600	
Error publicat	2,23	2,19	2,50	

Pràctica 1

Analitzeu les anomenades "fitxes tècniques" i les enquestes preelectorals publicades a la premsa una setmana abans de la diada electoral del 15 de març de 1992.

-Recupereu del disc de pràctiques el fitxer SONDEIG.FW3.

-Premeu **F5** des del marc per començar. Al full no apareixen les coordenades de les cel·les, per tant haureu de mirar la línia d'estat per saber a quina cel·la es troba el cursor.

-Com que l'anàlisi el fareu sobre una població gran (els electors de Catalunya), l'exactitud de la mida de la població influirà poc en els resultats. Deixeu, doncs, a la cel·la B1 el nombre "gran" que hi apareix.

-A la cel·la B3 es troba el nivell de confiança de l'enquesta. Entreu-hi, si no hi és, el 95.5 o el 95, segons el cas. El nivell de confiança d'un mostreig és un percentatge que mesura la probabilitat que els resultats reals (electorals en el nostre cas) de la població s'ajustin als resultats de la mostra amb un marge d'error que, tot seguit, calculareu. Aquest nivell de confiança quantifica la seguretat que hi ha d'encertar, sempre i quan el procés s'hagi fet correctament. En la majoria d'enquestes d'opinió s'agafa el 95,5% com a nivell de confiança, tot i que podríeu escollir d'entre quatre nivells més.

-Entreu a la cel·la B8 un 1. Apareixeran els rètols corresponents a l'opció triada.

-Entreu a la cel·la B10 la mida de la mostra (quantitat de persones entrevistades) que surt a la fitxa tècnica de l'enquesta (vegeu llista anterior). Apareixerà l'error màxim. A les fitxes tècniques, aquest error va precedit per quelcom

similar a "p=q=50%". Això significa que si el 50% dels entrevistats opinessin el mateix, l'error en extrapolar aquest percentatge a la població seria l'indicat.

-Entreu a la cel·la B12 el percentatge d'intenció de vot d'un determinat partit. Apareixerà l'error mostral corresponent al percentatge entrat, que serà més petit que l'error màxim esmentat a l'apartat anterior. Només en el cas que el percentatge entrat fos del 50%, els dos errors coincidirien. Dues files més avall trobareu l'interval de confiança.

-Canvieu el percentatge de la cel·la B12 i observeu-ne els canvis.

-Compareu els resultats reals amb els dels sondeigs preelectorals, tenint en compte els intervals de confiança per a cada partit i l'error màxim de la mostra.

-Varieu la mida de la mostra i el nivell de confiança. Traieu-ne conclusions.

Pràctica 2

Es tracta, ara, de fer servir l'opció 2 de l'aplicació SONDEIG.

Suposeu que voleu fer una enquesta al vostre centre. El problema inicial és saber quina mida ha de tenir la mostra per no cometre un error massa gran. És el procés invers a l'anterior.

-Entreu un 2 a la cel·la B8 per triar la segona opció. Observeu que els rètols han canviat.

-Suposeu que el nombre total d'alumnes (mida de la població) és 700. Entreu aquesta quantitat a la cel·la B1. Com que l'estudi és sobre els alumnes d'un centre d'ensenyament, la població no és massa gran. En aquest cas la mida de la població sí que influeix sobre els resultats finals.

-Entreu a la cel·la B10 el percentatge màxim d'error que podeu cometre, per exemple un 5. Apareix la mida de la mostra aconsellada per a no superar l'error entrat. S'ha de dir que aquest error és el màxim que s'obtindria suposant que el 50% dels enquestats estessin d'acord.

-Aneu canviant l'error i observeu-ne els resultats.

-Canvieu la mida de la població i treieu-ne conclusions.

ANNEX

Coeficient de determinació

Suposem que s'ha fet una regressió a partir d'n punts $(x_1,y_1),\ldots,(x_n,y_n)$. Sigui $F(x)$ la funció de regressió escollida a partir dels diferents models. Si es calculen els $F(x_i)$ no tenen perquè coincidir amb els y_i. Siguin $y'_i= F(x_i)$. Per a cada valor es pot calcular $e_i=y_i-y'_i$. Són els anomenats residus de la regressió. Aleshores, el coeficient de determinació ve donat per l'expressió:

$$R^2= 1 - \frac{\Sigma\, e_i{}^2}{\Sigma(y_i-\overline{y})^2}$$

on $\overline{y}$ és la mitjana dels valors y_i.

Es pot veure que $0\leq R^2\leq 1$. Com més proper a 1 sigui el valor de R^2, més s'ajusta el model de regressió escollit als punts inicials. Només en el cas de la regressió lineal és sempre certa la igualtat $r^2=R^2$, on r és el coeficient de correlació lineal o de Pearson.

Errors en un mostreig (per a SONDEIG)

L'expressió de l'error mostral, en tant per cent, que s'utilitza a l'aplicació SONDEIG és:

$$e = 100\cdot Z_\alpha\cdot\sqrt{p\cdot(1-p)/n}\cdot\sqrt{(N-n)/(N-1)} \quad \text{on}$$

p és la proporció observada a la mostra, en tant per 1

N és la mida de la població

n és la mida de la mostra

z_α és un coeficient que surt de l'aproximació de la binomial per la normal i que depèn del nivell de confiança escollit. En aquest cas es consideren els possibles valors següents:

Nivell de conf.	z_α
90	1,65
95	1,96
95,5	2
99	2,58
99,7	3

Quan **N** és gran es veu que l'últim factor de la fórmula és despreciable.

Es demostra que aquest error és màxim quan p=0.5 . És el cas identificat normalment com p=q=50%.

BIBLIOGRAFIA

CUADRAS, C. M. i d'altres. *Fundamentos de estadística. Aplicación a las Ciencias Humanas.* PPU. Barcelona 2a edició (1988)

DIVERSOS AUTORS. *L'estadística en el vostre món.* Schools Council publications i ICE de la UAB. Barcelona (1991)

FREEDMAN, D. i d'altres. *Stadistics.* WW Norton & Company. Londres 2a edició (1991).

LARSON, H.J. *Introducción a la teoría de probabilidades e inferencia estadística.* LIMUSA. México (1989).

NORTES, A. *Encuestas y precios.* SINTESIS. Madrid (1989).

THOMAS, J.J. *Introducción al análisis estadístico para economistas.* MARCOMBO. Barcelona (1986)